INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

New codecs : state-of-the-art techniques and prototypes

Jean-Christophe Mignot,
Laboratoire de l'Informatique du Parallélisme,

ENS Lyon, France

No 4074

Dec 2000

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--4074--FR+ENG

INRIA
RHÔNE-ALPES

New codecs : state-of-the-art techniques and prototypes

Jean-Christophe Mignot,
Laboratoire de l'Informatique du Parallélisme,
ENS Lyon, France

Thème 1 — Réseaux et systèmes
Projet ReMaP

Rapport de recherche n ° 4074 — Dec 2000 — 44 pages

Abstract: This paper presents a survey of the state-of-the-art techniques and prototypes for compressing videos. A particular interest is devoted to compression for streaming over the Internet. The basic principles are presented, the most important approaches are described.

(Résumé : tsvp)

This work was partially supported by the French Ministry of Culture within the SPIHD Project, contract n° 00.2.93.0173

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
Téléphone : 04 76 61 52 00 - International: +33 4 76 61 52 00
Télécopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

Les nouveaux codecs : un état de l'art

Résumé : Ce document présente les protocoles et outils de Distribution de contenus multimédia par flot continu (streaming) disponibles actuellement. Les principes de base sont présentés. Les approches les plus importantes (RTP, RTSP) sont décrites.

Ce travail a été partiellement financé par le ministère de l'Industrie dans le cadre du contrat SPIHD, convention n° 00.2.93.0173

1 Généralités

Le terme codec vient de la contraction des termes "codeur" et "décodeur" de la même manière que le terme modem est issu de la contraction des termes modulateur et démodulateur. D'un point de vue fonctionnel, un codec est une unité de traitement à une entrée et une sortie. Des exemples de codecs typiques sont les décodeurs audio, les encodeurs de vidéo ou encore les machines à effet. Le codec accepte un buffer en entrée, effectue certains traitements et place le résultat dans un buffer de sortie.

La plupart du temps, l'utilisation des codecs est transparente pour l'utilisateur. Soit ils sont directement intégrés dans les architectures de visionnage telles `Windows Media Player`, `Real Player 7` ou `Quick Time 4.0`, soit ils sont disponibles sous la forme de plug-in additionnels. Ainsi, `Quick Time 4.0` est livré avec les codecs `Sorenson Video`, `Cinepak` etc. alors que `Indeo 3.2` est disponible en téléchargement sur le site d'Intel.

La place du codec pour l'exemple d'une vidéoconférence sur le Web est illustré par la figure 1. Le codec est le dispositif logiciel et/ou matériel, qui encode les images vidéo avant de les émettre sur Internet. Les deux participants doivent donc posséder les mêmes codecs, ou des codecs compatibles. Dans le cas du visionnage d'une vidéo, le problème est le même : la vidéo a été encodée avec un certain codec et le fichier résultant stocké localement. Pour le visionner, il faut utiliser un logiciel utilisant le même codec que celui utilisé pour la compression ou un codec compatible. De la même, manière, si le fichier a été rendu disponible sur le Web, les utilisateurs distants devront disposer du même codec que celui utilisé pour la compression ou d'un codec compatible. Le rôle du codage est de compresser les données tout en respectant la qualité de restitution et en permettant la tolérance aux pannes. Le taux de compression appliqué peut être variable suivant les utilisations envisagées. On trouve des codecs adaptés à la bande passante du Web (par exemple 56 kb/s), à la bande passante des CD-ROM (typiquement 150 ko/s), à la bande passante des réseaux haut débit.

Un codec fonctionne dans un des deux modes suivants.

– Mode paquet. Le codec reçoit une image en entrée et la convertit en une image de sortie. Dans ce mode, le codec doit vider le buffer d'entrée pour générer celui de sortie. Ce mode est utile quand le codec accepte des

données de taille variable comme par exemple un simple codec de gain qui devrait appliquer un coefficient multiplicateur aux échantillons fournis en entrée. Ce mode est aussi utile quand le codec ne peut traiter que des données de taille fixe et prédéterminée et que le buffer d'entrée fournit des paquets de la taille de l'image. C'est par exemple le cas des décodeurs audio GSM qui reçoivent des paquets d'un dépaquettiseur RTP (Real Time Protocole, voir [5]), les décodent et placent le résultat dans le buffer de sortie.

– Mode flux. Le codec reçoit des morceaux de données en entrées et génère éventuellement des morceaux de données en sortie. Dans ce mode, le codec peut ne consommer qu'une partie du buffer d'entrée (le reste sera lus lors des appels suivants) et peut ne pas générer de buffer de sortie à chaque appel. Ce mode est utile pour les paquetiseurs de flux qui reçoivent un flux d'octets et divisent le flux en paquets (images) de sortie qui sont ensuite traités par un autre processus. Ce mode est aussi utile quand deux flux ayant des paquets de tailles différentes doivent être mixés.

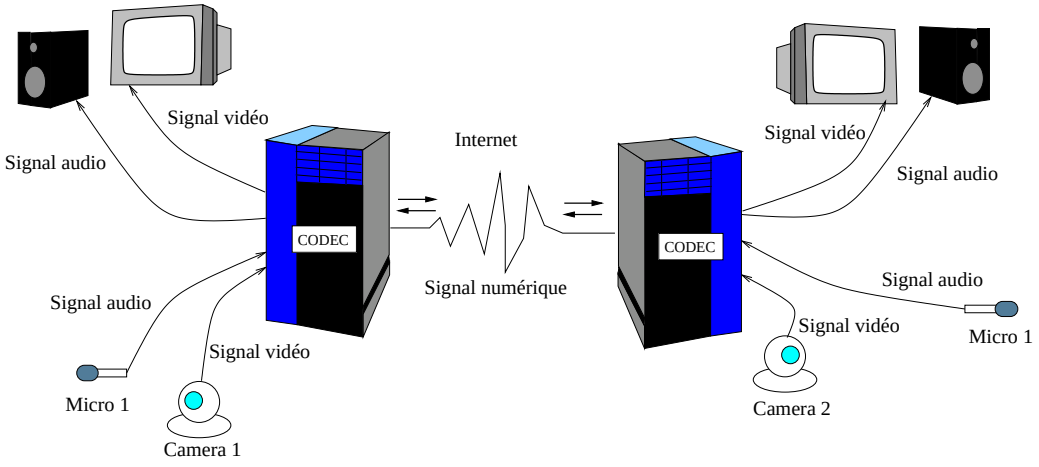

FIG. 1: La place du codec dans une vidéoconférence

Les taux de compression atteint par les codecs actuels sont très variables. Le taux de compression de 1 : 1 correspond par exemple aux codecs MJPEG

où la compression est nulle. Le film est constitué d'une suite d'images JPEG concaténées. Ce format est particulièrement adapté aux travaux de montage vidéo où les nombreux manipulations successives du fichier engendreraient un dégradation trop importante de la qualité de l'image si la compression était appliquée à chaque fois. Les codecs adaptés aux CD-ROM ont des taux de compression variant de 10 : 1 à 100 : 1. Citons par exemple MPEG-1, MPEG-2, Cinepak, Indeo, Video Interactive, Sorenson Video ou encore Apple Animation. Les codecs pour le Web doivent s'adapter à un débit de 10 à 100 fois moindres. Citons par exemple MPEG-4, Real Video et Sorenson Video.

La tolérance aux pertes caractérise la capacité du codec à reconstruire les paquets perdus ou à pallier cette perte. Cette capacité est d'actualité avec le développement rapide des communications mobiles à très bas débit et à taux de perte important. Les outils développé pour MPEG-4 concernant la tolérance aux pannes peuvent être divisés en 3 classes : resynchronisation, reconstruction de données et suppression d'erreur ([3]).

Concernant la vidéo, notons que la mesure de la qualité de l'image restituée par un codec est un problème ouvert. Si des méthodes permettent de mesurer le bruit ajouté à une image par une méthode de compression, aucune ne permet de refléter exactement l'impact de ce bruit sur la perception de la dégradation. En fait, le procédé le plus communément utilisé est d'avoir recours à des expérimentateurs et de leur demander d'attribuer des notes en fonction de la qualité perçue (voir [2]).

2 Quelques notions pour la compression de données

Nous décrirons dans cette section quelques méthodes de compression utilisées pour les données multimédia. La compression utilise trois types de redondance :

- la redondance spatiale, correspondant à la corrélation entre points adjacents d'une image ;
- la redondance spectrale, correspondant à la corrélation entre les différentes couleurs d'une image ;

– la redondance temporelle entre images successives dans une séquence vidéo.

On peut classifier les systèmes de compression d'image en systèmes avec ou sans perte d'information. Les systèmes de compression sans perte essayent de minimiser la bande passante de sortie sans distordre l'image de départ. Ces techniques sont utilisées quand la précision de l'image est essentielle ou quand de multiples séquences de compression/décompression doivent être appliquée (montage vidéo). Les systèmes de compression avec perte tentent d'obtenir la meilleure qualité d'image pour une bande passante donnée (ou inversement). Ce type de système est de loin le plus largement utilisé pour la distribution.

Outre la distinction par la perte d'information, on peut aussi classifier les systèmes de compression de données en codages entropiques et codages source. Les codages entropiques ne tiennent pas compte de la nature des informations compressées. Toutes les données sont traitées comme des séquence de bits sans sémantique. Les codages source exploitent la nature du signal original. Par exemple, si le signal d'origine est de la vidéo ou de l'audio, l'encodage de source utilise les caractéristiques de ce type de données pour obtenir un meilleur taux de compression.

Un système de compression typique est constitué d'un transformateur, d'un quantificateur et d'un codeur (voir la figure 2). Le transformateur est typiquement une application qui permet de rendre l'image plus facilement compressible. Par exemple, la transformation en cosinus discrète (Discrete Cosine Transform, appelée DCT dans la littérature anglophone) qui permet de réduire l'énergie du signal en un petit nombre de coefficients. Le quantificateur génère un nombre limité de symboles qui peuvent être utilisés pour la représentation de l'image compressée. Le codeur associe un code binaire à chaque symbole en sortie du quantificateur. Les codes peuvent être de taille fixe ou variable. Les codes de taille variable permettent de réduire la taille moyenne de la représentation binaire des symbole en affectant les codes les plus courts aux symboles les plus fréquents.

FIG. 2: Un système typique de compression d'image

2.1 Run-length encoding

Le principe de ce codage est simple : remplacer n occurrences successives d'une valeur donnée par le nombre d'occurrence et la valeur. Par exemple, la suite "3 3 3 3 6 6 12 5 5 5" sera remplacée par "4 3 2 6 1 12 3 5". Ce codage est souvent appelé codage RLE.

2.2 Les quantifications scalaire et vectorielle

La quantification vectorielle dans son sens le plus général est l'approximation d'un signal d'amplitude continue par un signal d'amplitude discrète. Elle peut être vue comme une application Q associant à chaque vecteur d'entrée $X = (x_j \in \mathbb{R}, j = 1..k)$ un vecteur $Y = (y_j \in \mathbb{R}, j = 1..k) = Q(X)$ choisi dans un dictionnaire $C = \{C_l \in \mathbb{R}^k, l = 1..N_c\}$ où N_c est le nombre d'éléments dans le dictionnaire. Pour $k = 1$, la quantification est dite scalaire. Un exemple de quantification scalaire est l'arrondi qui remplace un réel x par sa valeur arrondie n de $\mathbb{Z}$. Pour $k \neq 1$, la quantification est dite vectorielle.

La quantification vectorielle peut aussi être vue comme la combinaison d'un codeur et d'un décodeur. Le codeur prend en entrée un vecteur X, cherche dans le dictionnaire le vecteur qui lui ressemble le plus et transmet en sortie son indice j. Le décodeur reçoit en entrée l'indice j et génère le vecteur $Q(X) = C_j$ qui est un approximation de X.

Le principe de l'utilisation de la quantification vectorielle pour la compression d'image est illustré par la figure 3. Considérons un pixel donné en entrée $X = (x_1, x_2, x_3)$ où x_1, x_2 et x_3 sont ses valeur RGB, on cherche son correspondant C_j dans le dictionnaire. Seul est transmis au décodeur l'indice j. Le décodeur recevant l'indice retrouve le vecteur C_j dans le dictionnaire et le transmet au visionneur de l'image.

La recherche du correspondant dans le dictionnaire est basée sur une mesure de distorsion entre X et les différents C_l, notée d. Le vecteur C_j choisi est celui qui minimise la mesure de distorsion. Différentes mesures de distorsion ont été proposées dans la littérature. La mesure idéale doit évaluer la qualité subjective de l'image reconstituée. Faute de modèle fiable de qualité subjective, l'erreur quadratique moyenne est le plus souvent utilisée pour choisir $C_j = Q(X)$.

Si le dictionnaire contient tous les vecteurs de l'image à traiter, la compression sera due au fait que seuls les indices des vecteurs dans le dictionnaire sont

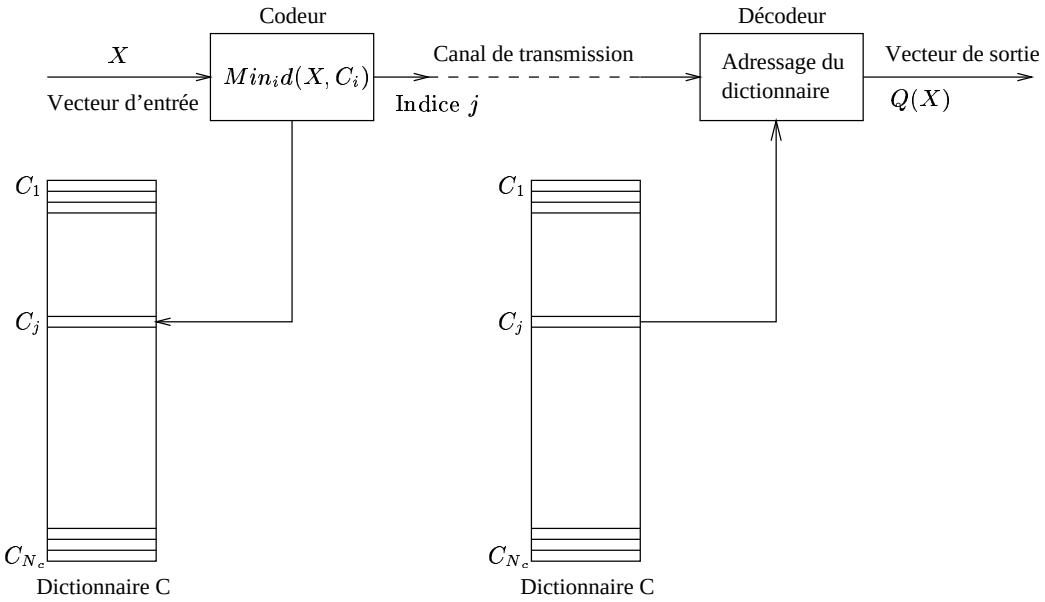

FIG. 3: Principe de la quantification vectorielle

transmis et non pas les vecteurs entiers. Si le dictionnaire est un sous-ensemble de l'ensemble des vecteurs de l'image, certains vecteurs du dictionnaires représenteront plusieurs vecteurs de l'image d'origine. La compression de la taille de l'image est alors due à la diminution de la taille du dictionnaire et à la transmission des indices seulement, comme précédemment.

2.3 La transformation en cosinus discrète

La transformation en cosinus discrète (Discrete Cosine Transform ou DCT pour les anglophones) a été rendue populaire de par son utilisation dans la norme JPEG. La DCT est très proche de la transformée de Fourier. Elle prend un ensemble de points d'un domaine spatial et les transforme en une représentation identique dans un domaine de fréquences. Cependant, dans le cas de la DCT, le signal d'entrée est un signal en 3 dimensions tracé suivant les axes X, Y et Z. Les axes X et Y sont les deux dimensions de l'image, l'axe des Z représente la valeur de chaque pixel en nuance de gris. L'image est vue

comme un signal tridimensionnel, c'est un représentation spatiale du signal. La DCT permet de convertir cette information spatiale en une information de fréquence, ou spectrale, dont les axes X et Y représentent les fréquences du signal en 2 dimensions. Les images couleur sont traitées comme 3 images différentes, une pour chacune des composantes R, G et B.

Pour une image de dimension $N \times N$, la transformation directe et la transformation inverse peuvent s'écrire respectivement ([4]) :

$$DCT(i,j) = \frac{1}{\sqrt{2N}} C(i)C(j) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} pix(x,y) cos(\frac{(2x+1)i\pi}{2N}) cos(\frac{(2y+1)j\pi}{2N}) \quad (1)$$

$$pix(x,y) = \frac{1}{\sqrt{2N}} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} C(i)C(j)DCT(i,j) cos(\frac{(2x+1)i\pi}{2N}) cos(\frac{(2y+1)j\pi}{2N}) \quad (2)$$

où $C(i) = \frac{1}{\sqrt{2}}$ si $i = 0$ et $C(i) = 1$ si $i > 0$, $pix(x,y)$ est la valeur du pixel de coordonnées (x,y), i et j les composantes en fréquence.

Le résultat obtenu pour une matrice d'entrée de taille $N \times N$ sera une nouvelle matrice de taille $N \times N$. Les valeurs sur la ligne 0 de cette matrice ont une composante en fréquence nulle dans une direction du signal, tous les éléments de la colonne 0 ont une composante en fréquence nulle dans l'autre direction. Lorsque les lignes et les colonnes s'éloignent de l'origine, les coefficients de la matrice transformée représentent des fréquences plus élevées, avec les plus hautes fréquences en position $N-1$. Le coefficient d'indice $(0,0)$ de la matrice transformée représente le « coefficient continu » de l'image.

Les images que nous visualisons sur ordinateur sont composées principalement d'informations basse fréquence (de gros objets, le ciel etc.). Les composantes en ligne et colonne 0 (dites composantes continues) transportent une information plus utile que les composantes de haute fréquence. Plus nous nous éloignons des composantes continues de l'image, plus les valeurs sont faibles et de peu d'importance pour la description de l'image.

L'utilisation de la DCT permet donc de distinguer les parties d'information dans le signal qui pourront être supprimées en affectant au minimum la qualité de l'image (les coefficients de haute fréquence). La représentation de l'image est concentrée dans les coefficients en haut à gauche de la matrice obtenue.

La complexité du calcul de la DCT étant en $O(N^2)$, elle n'est dans la pratique jamais appliquée sur une image entière. Il est préférable de décomposer

l'image en blocs de plus petite taille sur lesquels la DCT sera appliquée. Le comité JPEG a choisi des blocs de taille 8×8 rendant ainsi possible des compressions/décompressions rapides avec les technologies actuelles. On parle de codage par bloc. L'application de la DCT sur des blocs de taille plus grande pourrait donner des taux de compression plus élevés. Cependant des études ont montré que les connexions entre les pixels ont tendance à diminuer rapidement de sorte que même des pixels distants de 20 ou 30 positions sont très peu utiles pour les prédictions. Un bloc de 64×64 ne donnera donc pas une compression beaucoup plus élevée pour un temps de calcul nettement plus long.

De nombreuses techniques sont utilisées pour diminuer les temps de calcul de la DCT. Cette méthode étant utilisée dans de nombreuses normes, dont JPEG, l'engouement pour l'optimisation de la DCT est grand et la description de l'ensemble des optimisation sort du cadre de ce document. Signalons simplement que la DCT peut être implémentée sous la forme du produit de 3 matrices tel que décrit par l'équation 3.

$$DCT = C \times Pixels \times C^T \tag{3}$$

où $Pixels$ est la matrice des valeurs des pixels du bloc, la matrice C étant définie par :

$$\begin{array}{rcll} C_{i,j} & = & \frac{1}{\sqrt{N}} & \text{si } i = 0 \\ C_{i,j} & = & \sqrt{\frac{2}{N}} cos(\frac{(2j+1)i\pi}{2N}) & \text{si } i > 0 \end{array} \tag{4}$$

Le coût de la DCT devient alors de $2N$ additions et multiplications.

D'autres optimisations s'appuient sur la transformation de la DCT pour se ramener à des calculs entiers ou encore à des techniques de traitement numérique du signal appliquées à la transformée de Fourier.

3 La compression JPEG

La compression JPEG allie une DCT, une quantification, un codage RLE et un codage dit entropique. Après avoir appliqué la DCT sur un bloc de l'image, son occupation mémoire a augmenté. En effet, nous sommes passé d'une matrice 8×8 d'entiers à une matrice 8×8 de nombres flottants. La phase de quantification va permettre de se ramener à une matrice d'entiers

de petites valeurs réduisant ainsi le nombre de bits nécessaires au stockage de l'image. C'est une phase non conservative de JPEG.

La quantification consiste à réduire le nombre de bits nécessaires au stockage des nombres par la diminution de la précision des entiers. En plus de cette compression appliquée à tous les coefficients, il est possible de réduire davantage la précision des coefficients éloignés du coefficient $(0,0)$. Et plus on s'éloigne de ce coefficient plus la diminution de précision peut être grande puisqu'il correspondent à des fréquences de plus en plus élevées. Dans la pratique, on utilise des matrices de quantification où chaque coefficient correspond à la réduction de précision qui sera appliquée au coefficient de même position dans la matrice résultante de la DCT. Ces valeurs sont codées par des entiers entre 1 et 255 et sont souvent appelées «quantum». Aux éléments les plus importants sont associés les plus petits quantum, les valeurs grandissant avec l'éloignement de l'origine. Les formules de quantification et de déquantification sont données respectivement par :

$$Quantification(i,j) = [\frac{DCT(i,j)}{Quantum(i,j)}] \quad (5)$$

$$Dequantification(i,j) = Quantification(i,j) * Quantum(i,j) \quad (6)$$

où [...] représente l'arrondi à l'entier le plus proche.

Un grand nombre de possibilités s'offre pour la construction de la matrice de quantification. Le comité JPEG fournit un ensemble standard de valeurs de quantification pour les développeurs. Les effets destructeurs de la quantification peuvent être mesurés soit par l'erreur mathématique entre l'image d'entrée et celle de sortie, mais nous avons vu que celle-ci n'est pas suffisante, soit en faisant appel à l'oeil humain, procédé long et coûteux. Nelson propose dans [4] une matrice où les bandes diagonales sont incrémentées d'un pas donné suivant l'éloignement de l'origine, le pas servant de facteur de qualité. Pour un pas de 2, on obtient :

$$\begin{matrix} 3 & 5 & 7 & 9 & 11 & 13 & 15 & 17 \\ 5 & 7 & 9 & 11 & 13 & 15 & 17 & 19 \\ 7 & 9 & 11 & 13 & 15 & 17 & 19 & 21 \\ 9 & 11 & 13 & 15 & 17 & 19 & 21 & 23 \\ 11 & 13 & 15 & 17 & 19 & 21 & 23 & 25 \\ 13 & 15 & 17 & 19 & 21 & 23 & 25 & 27 \\ 15 & 17 & 19 & 21 & 23 & 25 & 27 & 29 \\ 17 & 19 & 21 & 23 & 25 & 27 & 29 & 31 \\ 19 & 21 & 23 & 25 & 27 & 29 & 31 & 33 \end{matrix} \tag{7}$$

A titre d'exemple, considérons la matrice résultante d'une DCT représentée en (8).

$$\begin{pmatrix} 92 & 3 & -9 & -7 & 3 & -1 & 0 & 2 \\ -39 & -58 & 12 & 17 & -2 & 2 & 4 & 2 \\ -84 & 62 & 1 & -18 & 3 & 4 & -5 & 5 \\ -52 & -36 & -10 & 14 & -10 & 4 & -2 & 0 \\ -86 & -40 & 49 & -7 & 17 & -6 & -2 & 5 \\ -62 & 65 & -12 & -2 & 3 & -8 & -2 & 0 \\ -17 & 14 & -36 & 17 & -11 & 3 & 3 & -1 \\ -54 & 32 & -9 & -9 & 22 & 0 & 1 & 3 \end{pmatrix} \tag{8}$$

L'application de la matrice de quantification présentée en (7) à la matrice de sortie de DCT représentée en (8) donne la matrice présentée en (9). On notera le nombre important de 0 de plus en plus nombreux quand on s'éloigne des composantes continues.

$$\begin{pmatrix} 90 & 0 & -7 & 0 & 0 & 0 & 0 & 0 \\ -35 & -56 & 9 & 11 & 0 & 0 & 0 & 0 \\ -84 & 54 & 0 & -13 & 0 & 0 & 0 & 0 \\ -45 & -33 & 0 & 0 & 0 & 0 & 0 & 0 \\ -77 & -39 & 45 & 0 & 0 & 0 & 0 & 0 \\ -52 & 60 & 0 & 0 & 0 & 0 & 0 & 0 \\ -15 & 0 & -19 & 0 & 0 & 0 & 0 & 0 \\ -51 & 19 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{9}$$

L'étape finale de la norme JPEG est le codage. Il utilise le fait que la matrice quantifiée contient beaucoup de 0 et que ceux-ci sont principalement regroupés en bas à droite. L'idée est d'utiliser un codage du type RLE et de parcourir la matrice en «zigzag» dans l'espoir d'optimiser la longueur des suites de 0. Le parcours en «zigzag» est illustré par la figure 4.

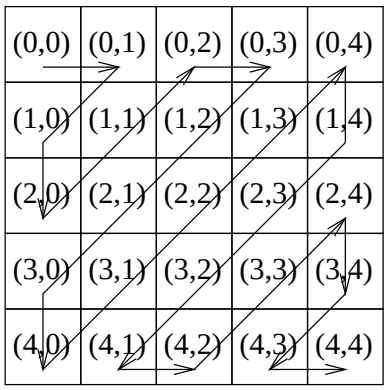

FIG. 4: Le parcours en zigzag d'une matrice

La dernière phase de compression de la norme JPEG est appelée «codage de l'entropie». Elle consiste à réduire le nombre de bits nécessaires à la représentation des valeurs de la matrice au strict minimum. Le principe est simple : la représentation binaire classique est remplacée par une représentation du type $<$ *nombre de bits, valeurs des bits*$>$ où *nombre de bits* représente le nombre de bits nécessaires pour coder la valeur en binaire et où *valeurs des bits* est la chaîne de bit codant la valeur. Le gain est particulièrement intéressant surtout si les matrices après traitement on la forme supposée, à savoir un grand nombre de petites valeurs et de rares grandes valeurs.

Finalement, on obtient un fichier contenant un ensemble de triplets représentés par la figure 5.

4 La compression MPEG2

Définie en 1994, la compression MPEG2 est très utilisée depuis 1997 avec l'adoption de ce standard pour les DVB (Digital Video Broadcasting, pour la

Nombre de zéros consécutifs avant la valeur courante	Nombre de bits nécessaires pour coder la valeur	Amplitude du coefficient de la DCT

FIG. 5: Un triplet d'un fichier au format JPEG

diffusion numérique de la télévision) et les DVD (Digital Versatile Disc, pour la diffusion de vidéo sur disque compact). Une séquence vidéo compressée en MPEG-2 occupe environ 30% d'espace en moins que la même séquence compressée en M-JPEG.

Les principales qualités de MPEG par rapport aux compressions classiques sont au nombre de trois.

- MPEG est spécialement optimisé pour la vidéo : un seul type d'espace des couleurs est autorisé ; seuls un certain nombre de résolutions et de taux de compression sont permis ; enfin, des mécanismes de gestion du flux audio sont intégrés.
- MPEG garantit que le taux de transfert demandé est celui qui sera obtenu, facilitant ainsi la gestion des admissions des clients par les serveurs.
- MPEG se sert de 2 caractéristiques fondamentales des vidéos : le haut degré de redondance entre deux images successives d'une vidéo et la nature généralement prédictible des mouvements.

Deux types de compression sont utilisés dans le codage MPEG : la compression spatiale et la compression temporelle. Elles sont décrites dans les sections qui suivent. Le but de ces sections est d'indiquer les principes de base utilisés. Le lecteur intéressé pourra se référer par exemple à `http://www.mpeg.org/` pour des informations détaillées incluant les codes source pour les encodeurs, décodeur et visionneurs et à `http://www.cselt.it/mpeg/` pour la page officielle du MPEG Group.

4.1 Compression spatiale

La compression spatiale est similaire à celle utilisée par la norme JPEG. Elle utilise la similarité entre des pixels adjacents sur une surface unie et tient

compte des fréquences spatiales dominantes. Il est possible de coder une séquence vidéo comme une succession d'images JPEG, on parle alors de codage M-JPEG ou Motion-JPEG. La compression temporelle n'est pas utilisée, les taux de compression sont donc moins bons que quand les 2 types de codages sont utilisés. Cependant, le découpage de la vidéo image par image est facilité et les dégradations de qualité par rapport à l'original sont moindres. Ce codage est donc principalement réservé aux applications demandant une dégradation minimale et un découpage aisé comme les applications de montage vidéo.

4.2 Compression temporelle

En MPEG, la redondance temporelle est d'abord réduite en utilisant les similitudes entre deux images successives. La plus grande partie de l'image courante est créée (on trouve parfois le terme «prédictée») en utilisant l'information de l'image déjà émise. En utilisant cette technique, il suffit de transmettre une image de différence qui élimine les différences entre l'image actuelle et l'image de prédiction. L'image de différence est ensuite soumise à une compression spatiale. Le décodeur inverse le codage en ajoutant l'image de différence à l'image précédente pour obtenir l'image suivante. De la même manière, la différence peut être définie par rapport à une image à venir. Le codage bidirectionnel réduit considérablement la quantité de données de différence nécessaire à l'amélioration du degré de prédiction. Un codeur intelligent devra essayer les trois stratégies de codage et sélectionner celle qui nécessite la transmission d'un minimum de données.

Cette technique pose quelques problèmes. Comme seules les différences sont transmises, le décodage ne peut commencer n'importe où, ce qui peut-être problématique après une commutation d'un flux de données vers un autre (changement de canal). De plus, si une erreur de transmission est intervenue, elle persistera indéfiniment dans l'image. La solution à ce problème consiste à utiliser un système qui n'est pas uniquement différentiel. Périodiquement, des images complètes sont transmises, elles sont appelées images intra-codées souvent dénotées images I, elles sont obtenues en n'appliquant qu'une compression spatiale à l'image d'origine. Si une commutation de flux intervient ou si une erreur se produit, on retrouve un décodage correct à l'image I suivante.

4.3 Compensation de mouvement

Le mouvement réduit la similitude entre deux images et augmente la quantité de données nécessaires à la création de l'image de différence. La compensation de mouvement est utilisée pour accroître cette similitude. La figure 6 en montre le principe.

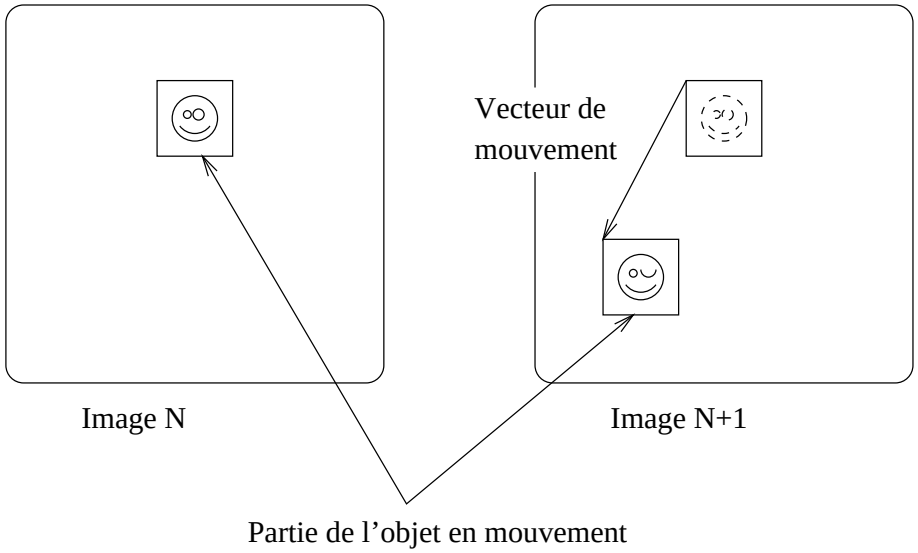

FIG. 6: Le principe de la compensation de mouvement

Quand un objet se déplace sur un écran, il apparaît à un endroit différent, mais il ne change pas beaucoup d'aspect. La différence d'image peut être réduite en mesurant le déplacement lors du codage. Ce déplacement est transmis au décodeur sous la forme d'un vecteur. Le décodeur utilise ce vecteur pour décaler une partie de l'image précédente vers l'emplacement approprié dans la nouvelle image. Un vecteur concerne le déplacement d'une zone entière de l'image appelée «macrobloc».

4.4 Codage bi-directionnel

Les images prédites à partir d'une image antérieure sont appelées images P. L'image qui sert de base à la prédiction peut être de type I ou P. Les données

d'une image P sont constituées de vecteurs décrivant où chaque macrobloc doit être pris dans l'image précédente et des coefficients non transformés décrivant la correction ou les données de différence à ajouter à ce macrobloc. Expérimentalement, on observe que les images P comportent pratiquement la moitié des données d'une image I.

Les images B sont prédictées bidirectionnellement à partir d'images antérieures ou postérieures et de type I ou P. Les données des images de type B consistent en vecteurs décrivant l'endroit où les données doivent être prises dans les images antérieures ou postérieures. Elles contiennent également les coefficients de transformée fournissant la correction.

On appelle GOP ou groupe d'images (GOP = Group Of Pictures) une séquence d'images commençant par une image I, suivie d'images P espacées par des images de type B. La fin du GOP se situe à la dernière image précédant immédiatement une nouvelle image I. La longueur d'un GOP est variable, mais les valeurs les plus courantes se situent entre 12 et 15. En fait, si les données d'une image B doivent être utilisées pour construire une image ultérieure, ces données doivent rester disponibles dans le décodeur. Par conséquent, le codage bidirectionnel implique que les données soient extraites de la séquence et provisoirement sauvegardées ce qui limite la taille maximale d'un GOP.

A cause des images bidirectionnelles, l'ordre de succession de ces images est différent à l'intérieur d'un groupe d'images (GOP) selon que l'on considère la transmission des images vers le récepteur, ou l'affichage des images sur ce même récepteur. En effet, on doit transmettre d'abord I et P pour qu'ensuite l'interpolation des images B soit possible. L'interpolation des images B suivantes est réalisée à partir de deux images P, et ainsi de suite jusqu'à l'image I suivante du groupe d'images suivant. La figure 7 illustre les changements réalisés. Les images P sont émises avant les images B. Les dernières images B du GOP ne peuvent être transmises qu'après la première image I du GOP suivant puisqu'elles ont besoin de son contenu pour être décodées bidirectionnellement.

4.5 Profils et niveaux

MPEG peut s'utiliser dans diverses applications avec des performances et des complexités diverses. Avec les outils de codage définis dans MPEG, il existe des milliers de combinaisons possibles. Dans un but de simplification, MPEG

# image initiale	0	1	2	3	4	5	6	7	8	9	10	11	12	
GOP compression	I	B	B	P	B	B	P	B	B	P	B	B	I	...
GOP affichage	I	P	B	B	P	B	B	P	B	B	I	B	B	...

FIG. 7: Le schéma de transmission des différents types d'image

est divisé en Profils, chaque profil étant lui-même subdivisé en Niveaux. Un profil constitue à la base la palette des caractéristiques d'un codage d'une certaine complexité. Un niveau est en fait un paramètre définissant par exemple la taille de l'image ou le débit du flux de bits. Il existe en principe 24 combinaisons possibles (6 profils et 4 niveaux) mais toutes n'ont pas été définies. Un décodeur MPEG possédant un profil et un niveau donné doit pouvoir décoder les signaux émanant d'un profil et d'un niveau inférieurs. Les profils sont appelés profil simple, principal, 4 :2 :2, SNR, spatial ou encore profil haut. Les niveaux sont appelés bas, principal, haut-1440 et haut.

Parmi les principaux profils, citons le profil le plus élémentaire appelé «profil simple». Il ne comporte pas de codage bidirectionnel, seules des images I et P sont générées par le codeur. Les délais de codage et de décodage sont réduits et le matériel nécessaire est simple. Un seul niveau est défini, le niveau «Principal». Le Profil Principal (Main Profile) est conçu pour une vaste gamme d'utilisations. Le niveau Bas utilise un signal d'entrée à faible résolution ne possédant que 352 pixels par ligne. La plupart des applications de diffusion nécessite le Profil principal au Niveau Principal (Main Profile at Main Level = MP@ML), appellation du MPEG utilisé en télévision standard. Le niveau Haut-1440 est un système à haute définition qui double la définition par rapport au niveau principal. Le niveau Haut double non seulement la définition horizontale, mais maintient cette résolution pour le format 16 :9 en portant à 1920 le nombre d'échantillons horizontaux.

5 La compression MPEG4

La compression MPEG4 est récente puisque la version 2 de la norme MPEG4 (compatible avec MPEG1 et MPEG2) a été finalisée fin 1999 ([3]).

Un des rares visionneurs permettant à l'heure actuelle de visualiser des vidéos suivant ce standard est la version 7 du Windows Media Player de Microsoft.

L'objectif premier de la norme MPEG4 était de succéder aux normes MPEG1 pour la compression et le transfert audio-vidéo et MPEG2 pour la télévision numérique. Mais lors de son élaboration, le champ des applications et des fonctionnalités a largement dépassé le cadre d'une simple évolution. MPEG4 est une norme vaste et très novatrice tant au niveau conceptuel que des applications visées.

MPEG4 réunit trois mondes : l'informatique, les télécommunications et la télévision. Elle est le résultat d'un effort international regroupant des centaines d'ingénieurs et de chercheurs du monde entier et de divers milieux : universités, centres de recherche, grands groupes informatiques, de télécommunication et autres grands groupes industriels.

MPEG4 marie la télévision digitale, les applications graphiques interactives (images de synthèse) et le multimédia interactif (WWW, la distribution et l'accès au contenu). La norme fournit les éléments technologiques standards pour permettre l'intégration de la production, de la distribution et de l'accès à ces trois domaines.

Dans la suite de cette section, nous décrirons dans un premier temps les principales caractéristiques de la normes MPEG4. Tous les aspects ne seront pas abordés. La norme est vaste, elle aborde les aspects réseau, le format des fichiers, les aspects audio et vidéo, la gestion des erreurs, la définition de profils, les procédures d'évaluation de la qualité du rendu et la gestion des droits d'auteurs. Dans un deuxième temps, nous nous intéresserons aux aspects visuels de la norme puisqu'ils remplissent un rôle majeur : diminuer la bande passante nécessaire au visionnage d'un film. Notons d'ailleurs qu'aucun système actuel ne couvre l'ensemble de la norme, les visionneurs se contentant d'implémenter la partie visuelle (et surtout la sous-partie vidéo), puisqu'elle apporte le plus de valeur ajoutée.

5.1 Introduction

Le standard MPEG4 fournit un ensemble de technologies satisfaisant le besoin des auteurs, des fournisseurs et des utilisateurs.

Pour les auteurs, MPEG4 permet la production de séquences de réutilisabilité et flexibilité nettement plus grandes qu'aujourd'hui en utilisant les technologies de télévision numérique, graphique animé ou de pages Web. En outre, les auteurs ont la possibilité de gérer et de protéger leurs droits d'auteurs.

Pour les fournisseurs d'accès Internet, MPEG4 offre une information transparente qu'ils pourront aisément interpréter et traduire en signaux adaptés à leurs réseaux grâce à des outils standards. Une exception à cette règle est la qualité de service qui est exprimée à l'aide de descripteurs génériques de qualité de service pour différents média MPEG4. La traduction des différents ensembles de paramètres de qualité de service pour chaque média vers la qualité de service du réseau est laissée à la charge des prestataires de service.

Pour les utilisateurs, MPEG4 permet un plus haut niveau d'interaction avec le contenu, dans les limites fixées par le producteur. La norme permet aussi l'accès du multimédia aux nouveaux réseaux, parmi lesquels les réseaux bas débit et les réseaux pour mobiles.

Pour tous ces intervenants, MPEG4 essaie d'éviter la profusion de formats et de visionneurs propriétaires ou incompatibles. Ce but est atteint par la fourniture de solutions standards pour :

- la représentation des «objets médias» (unité sonore, visuelle et audiovisuelle). Ces objets médias peuvent être naturels ou synthétiques (i.e. enregistrées par un appareil photo, un microphone ou générées par un ordinateur),
- la composition des objets entre eux pour créer les objets composés formant une scène audiovisuelle,
- le multiplexage et la synchronisation des données associées aux objets média afin qu'ils puissent être transportés sur des réseaux en fournissant une qualité de service adapté à la nature de l'objet,
- l'interaction entre l'utilisateur et la scène audiovisuelle générée du coté utilisateur.

La suite de cette section illustre les fonctionnalités de MPEG4 décrites ci-dessus en utilisant la scène audiovisuelle représentée par la figure 8.

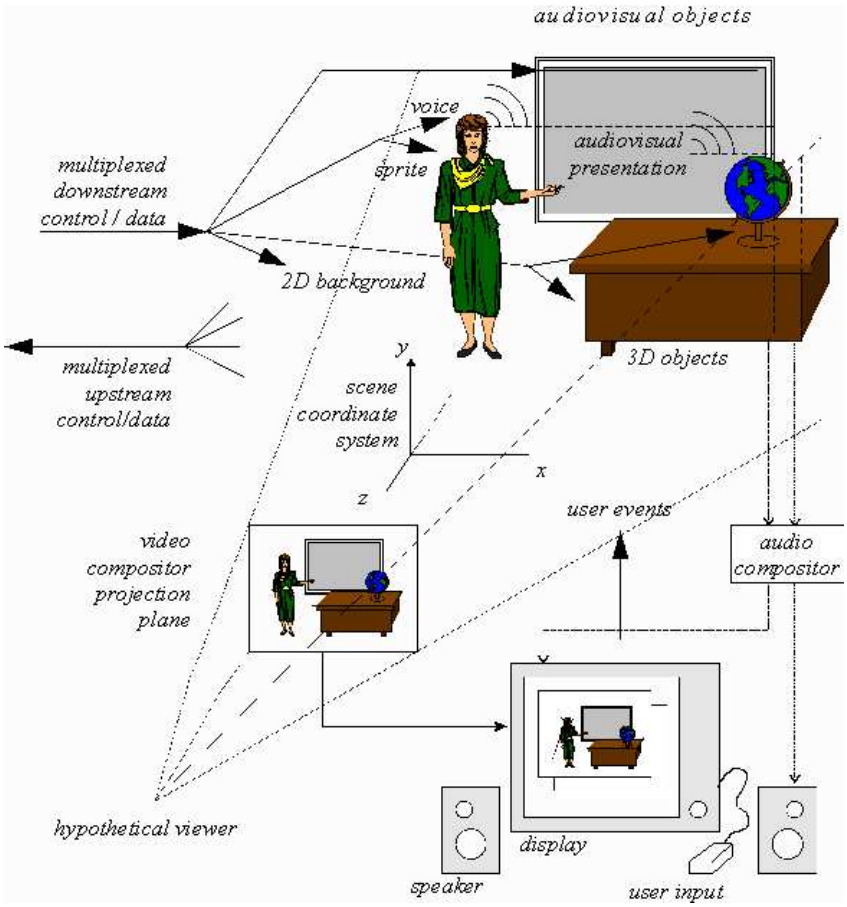

FIG. 8: Exemple de scène suivant la norme MPEG4

5.1.1 La représentation codée des objets média

Les scènes audiovisuelles MPEG4 sont composées de plusieurs objets média organisés de manière hiérarchique. Les feuilles de la hiérarchie sont les objets média élémentaires tels les images fixes (par exemple le fond de la vidéo), les objets vidéo (comme une personne qui parle *sans* le fond), les objets audio (la voix associée à la personne), etc.

MPEG4 standardise un certain nombre d'objets média élémentaires qu'ils soient naturels ou synthétiques, en 2 ou 3 dimensions. En plus de ces objets, MPEG4 définit la représentation codée pour des objets tels les textes et graphiques, les visages synthétiques parlants et leurs textes associés utilisés pour synthétiser la voix et l'animation du visage, le son synthétique.

Un objet média dans sa forme codée consiste en des éléments descriptifs qui permettent de gérer l'objet dans un scène audiovisuelle et éventuellement les flux de données associés. Il est important de noter que dans sa forme codée chaque objet média peut être représenté indépendamment de son environnement ou du fond.

La représentation codée des objets média est aussi efficace que possible tout en tenant compte des spécificités désirées pour l'objet. Des exemples de telles spécificités sont la robustesse aux erreurs, la facilité d'extraction et d'édition de l'objet ou d'avoir un objet redimensionnable.

5.1.2 La composition des objets média

La figure 8 explique la manière dont une scène audiovisuelle est décrite comme composée d'objets individuels. La figure contient des objets média composés qui regroupent des objets média élémentaires. La structure hiérarchique de la scène est représentée par la figure 9. Les objets média élémentaires correspondent aux feuilles dans l'arbre de description tandis que les objets composés correspondent à des sous-arbres. Par exemple, l'objet visuel qui correspond à la personne qui parle et la voix associée sont liés pour former un nouvel objet média composé contenant à la fois les composant oraux et visuels de la personne qui parle.

De tels regroupements permettent aux auteurs de construire des scènes complexes et aux consommateurs de manipuler des entités compréhensibles (une personne parlante plutôt qu'une image de personne avec une bande son).

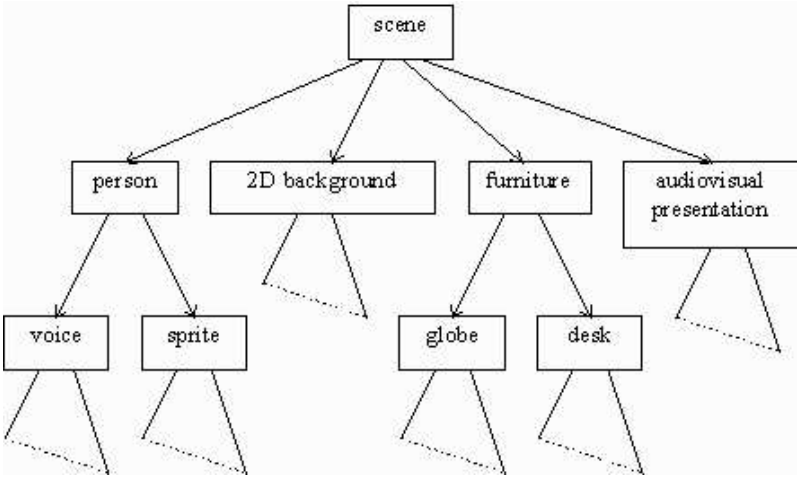

FIG. 9: Schéma de structure d'une scène suivant la norme MPEG4

Plus généralement, MPEG4 fournit des moyens standards pour décrire une scène, permettant par exemple de :

- de placer un objet n'importe où dans un système de coordonnées,
- d'effectuer des transformations géométriques ou acoustiques sur un objet,
- de grouper des éléments média simples pour former un composant média complexe,
- de modifier les attributs d'un objet à partir d'un flux de données entrant,
- de changer interactivement, la vue et l'écoute d'une scène.

La manière dont les scènes sont décrites est largement inspirée des concepts issus de VRML (Virtual Reality Modeling Language) tant du point de vue de la structure que des fonctionnalités des nœuds de composition d'objets et les étend pour permettre toutes les fonctionnalités précitées.

5.1.3 Description et synchronisation des flux données pour les objets média

Les objets média peuvent avoir besoin d'un flux de données transporté par un ou plusieurs flux élémentaires. Un descripteur d'objet identifie tous les flux

associés à un objet média. Il permet de gérer les objets codés hiérarchiquement, l'association de méta-informations sur le contenu (appelé «information sur le contenu de l'objet») et les droits de propriété intellectuelle associés.

Chaque flux est caractérisé par un ensemble de descripteurs pour les informations de configuration par exemple pour déterminer les ressources en décodeurs nécessaires ou la précision des informations de temps. De plus, les descripteurs peuvent contenir des informations relatives à la qualité de service nécessaire pour la transmission du flux (comme le débit maximum, la priorité etc.).

La synchronisation des flux élémentaires est réalisée par un estampillage des unités d'accès individuelles à l'intérieur des flux élémentaires. La couche de synchronisation gère l'identification de telles unités d'accès ainsi que l'estampille temporelle. Indépendamment du type d'objet média, cette couche permet l'identification du type des unités d'accès (par exemple des images, des commandes de description de scène) dans les flux élémentaires, la détermination de la base de temps pour des objets média ou des description de scène et permet la synchronisation entre eux. La syntaxe de cette couche est largement configurable permettant une utilisation par de nombreux systèmes.

5.1.4 Diffusion des flux de données

La diffusion synchronisée de flux d'informations de la source à la destination, utilisant les différentes qualités de service disponibles sur le réseau, est assurée par la couche de synchronisation (décrite précédemment) et une couche de diffusion contenant un multiplexeur à deux couches. Le principe de fonctionnement est illustré par la figure 10. Le terme TransMux représente une abstraction générique pour n'importe quel schéma de transport multiplexé (de l'anglais Transport Multiplexing) robuste aux erreurs et issu de l'organe de communication. Le terme FlexMux (de l'anglais Flexible Multiplexing) représente une couche de multiplexage de niveau supérieur. Cette couche ne gère pas les erreurs puisqu'elles sont gérées par la couche TransMux. Un flot TransMux peut comporter plusieurs flots FlexMux. Un flot FlexMux peut comporter plusieurs flots élémentaires. Le terme DMIF vient de l'anglais Delivery Multimedia Integration Framework. Cette couche offre une interface unifiée pour l'accès aux réseaux et fichiers. Elle permet aux développeurs de n'écrire qu'une

application pour différentes sources de données. La couche Sync Layer est la couche de synchronisation des flux élémentaires.

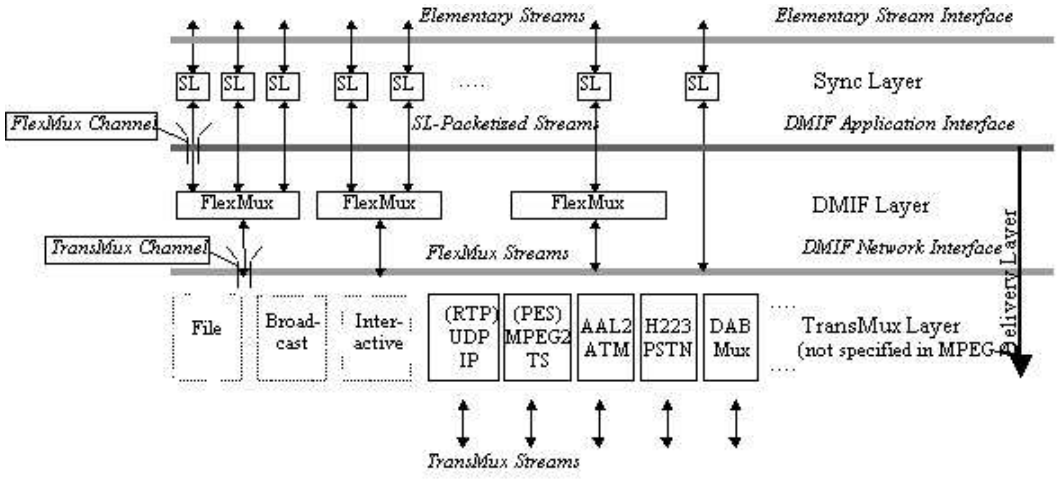

FIG. 10: Le modèle en couche du système de la norme MPEG4

La première couche de multiplexage est gérée suivant les spécifications du DMIF (Delivery Multimedia Integration Framework), qui est une partie du standard MPEG4. Ce multiplexage peut être intégré dans un outil FlexMux qui permet le regroupement de différents flots élémentaires avec un surcoût minimal. Dans cette couche le multiplexage peut être utilisé pour par exemple regrouper des flots élémentaires nécessitant la même qualité de service, réduire le nombre de connexions au réseau ou encore le délai du départ à l'arrivée.

La couche TransMux représente la couche qui offre les services de transport correspondant à la qualité de service demandée. Seule l'interface de cette couche est spécifiée par MPEG4. La gestion des paquets de données et des signaux de contrôle est à la charge des responsables des protocoles de transport. De nombreux protocoles de transport peuvent être utilisés comme (RTP)/UDP/IP, (AAL5)/ATM etc. Ils représentent autant d'instance d'un TransMux. Le choix du TransMux est laissé à l'utilisateur final ou au prestataire de service, permettant à MPEG4 d'être utilisé sur une grande variété d'environnements.

L'utilisation de l'outil de multiplexage FlexMux est optionnelle et cette couche peut être vide si le TransMux sous-jacent fournit toutes les fonctionnalités requises. La couche de synchronisation est, elle, toujours présente.

Au vu de la figure 10, il est possible de :

- identifier les unités d'accès, les estampilles temporelles, l'horloge de référence et les pertes de données,
- multiplexer les données de différents flux élémentaires en flux FlexMux,
- envoyer des informations de contrôle pour indiquer les qualités de service pour les flux élémentaires et les flux TransMux, traduire ces qualité de service pour le réseau réellement utilisé, associer les flux élémentaires aux objets média et gérer la correspondance entre les flux élémentaires et les flux FlexMux et TransMux.

5.1.5 Interaction avec les objets média

En général, l'utilisateur visualise une scène qui est composée d'après le design de l'auteur. Cependant, suivant les degrés de liberté laissés par l'auteur, l'utilisateur a la possibilité d'interagir avec la scène. Parmi les opérations possibles, citons :

- le changement de point de vue, c'est-à-dire le déplacement dans la scène,
- le déplacement d'objets dans la scène,
- le déclenchement d'événements par des clics de souris comme par exemple pour démarrer ou stopper un flux vidéo,
- la sélection d'un langage quand plusieurs langages sont disponibles.

5.1.6 Gestion et identification de la propriété intellectuelle

Il a semblé important aux rédacteurs de la norme de donner la possibilité d'identifier la propriété intellectuelle pour les objets média. MPEG4 incorpore l'identification de la propriété intellectuelle en stockant un identificateur unique donné par un système international de numérotation (comme ISAN : Int. Standard Audiovisual Number ou ISRC : Int. Standard Recording Code). Ces nombres peuvent être utilisés pour identifier le détenteur des droits des objets média. Comme tous les contenus ne seront pas identifiés par un nombre,

MPEG4 fournit la possibilité d'identifier la propriété intellectuelle grâce à des paires de clés du style «compositeur»/«Jean Dupond». MPEG4 offre aussi une interface standardisée qui est couplée avec la couche System pour permettre l'insertion d'outils de contrôle d'accès à la propriété intellectuelle. Pour de plus amples renseignements, le lecteur intéressé pourra se référer à [1].

5.1.7 Vue générale

L'ensemble des spécifications abordées précédemment s'organise en un système MPEG4 illustré par la figure 11. Les flux émis depuis le réseau (ou une unité de stockage) comme des flux TransMux sont démultiplexés en flux FlexMux et transmis au démultiplexeur FlexMux approprié qui retrouvera les flux élémentaires (souvent désignés par ES pour Elementary Streams) contenus. Les flux élémentaires sont analysés et transmis aux décodeurs appropriés. Le décodage retrouve les données des objets depuis leur forme codée et applique les transformations nécessaires pour reconstruire les objet AV (Audio-Video) prêts pour le rendu sur l'organe de sortie. Les objets AV reconstruits sont mis à la disposition de la couche de composition pour une utilisation potentielle pour la phase de rendu. Les objets décodés et les informations de la description de la scène sont utilisés pour reconstruire la scène définie par l'auteur. L'utilisateur peut, dans des limites définies par l'auteur, interagir avec la scène présentée. Finalement la gestion et la protection des droits d'auteur sont appliquées et les informations sur le contenu des objets (OCI pour Object Content Information) interprétées.

5.2 Les aspects visuels de la norme MPEG4

5.2.1 La description de scènes

Comme nous l'avons vu précédemment, une scène est décrite par un arbre (figure 9). Lors de la diffusion cet arbre est codé et transmis en même temps que les objets média. Inspiré de VRML, MPEG a développé un langage pour la description de scène appelé BIFS pour BInary Formats for Scenes.

Pour faciliter le développement des outils de création, de manipulation et d'interaction, les descriptions de scène sont codées indépendamment des flux relatifs aux objets média élémentaires. Une attention spéciale a été portée

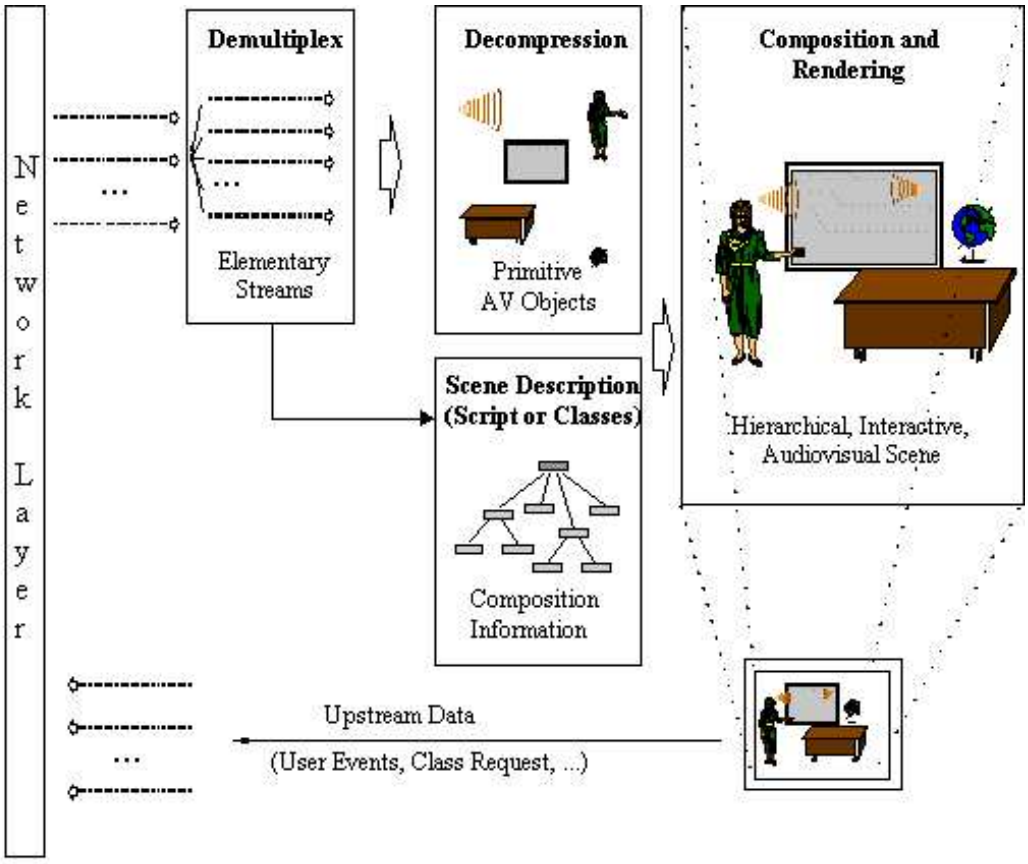

FIG. 11: Un système MPEG4 complet

à l'identification des paramètres appartenant à une description de scène en différenciant les paramètres utilisés pour améliorer l'efficacité du codage des objets (comme les vecteur de compensation de mouvement) de ceux qui sont utilisés comme des modificateurs des objets (comme la position d'un objet dans une scène). Pour permettre la modification de l'ensemble de ces derniers paramètres sans avoir à décoder les objets média élémentaires eux-mêmes, ces paramètres ont été placés dans la description de la scène et non dans les objets.

La liste suivante donne quelques exemples des paramètres contenus dans une description de scène.

- *Le regroupement des objets.* Une scène MPEG4 suit une structure hiérarchique qui peut être représentée comme un graphe acyclique (figure 9). Chaque feuille du graphe représente un objet média. La structure de l'arbre n'est pas nécessairement statique ; les attributs des feuilles (comme les paramètres de positionnement) peuvent être changées. On peut aussi envisager d'en ajouter, des nœuds peuvent être supprimés, remplacés ou même ajoutés.
- *Le positionnement des objets dans l'espace et le temps.* Dans le modèle MPEG4, les objets audiovisuels ont une composante spatiale et une composante temporelle. Chaque objet média a un système de coordonnées local. Un système de coordonnées local est un système dans lequel l'objet a une localisation spatio-temporelle et une échelle fixes. Ce système est utilisé pour manipuler l'objet dans le temps et l'espace. Les objets média sont positionnés dans la scène par la spécification d'un changement de repère du système de coordonnées local à l'objet vers un système global de coordonnées défini par un ou plusieurs nœuds de description de scène parents dans l'arbre.
- *La sélection des attributs de valeur.* Les objets média et les nœuds de la description de scène contiennent un ensemble de paramètres pour la couche de composition grâce auxquels une partie de leur comportement peut être contrôlée. On trouve par exemple la hauteur d'une note ou encore la couleur d'un objet synthétique.

5.2.2 Codage des objets visuels

La norme MPEG4 définit des standards tant pour les objets vidéo naturels que pour les objets synthétiques.

Les applications du standard vidéo MPEG4 Le standard MPEG4 vidéo offre une technologie qui couvre un large champ d'applications anciennes ou récentes. Le codage pour faibles débits et tolérant aux erreurs permet la transmission sur des canaux à faible débit sans fil comme ceux des téléphones portables ou des communications spatiales. Il peut être aussi utile dans le domaine de la télésurveillance ou les débits sont faibles ou très variables. Pour les hauts débits, des outils sont disponibles qui permettent la transmission et le stockage de vidéos de haute qualité acceptables pour les studios de montage ou autres applications de création de contenus. Il est possible que le standard supporte des débits bien au-delà de ceux de MPEG-2.

Un des domaines d'applications majeure est la vidéo interactive sur le Web. Des programmes permettant la diffusion en direct de vidéos MPEG4 sur le Web ont déjà été réalisés. L'orientation objet de MPEG4 ouvre de nombreux horizons. Les outils de codage de formes en niveau de gris ou binaire permettent de composer des objets vidéos quelconques avec des graphiques et des textes. Ces possibilités devraient permettre le développement de nouveaux types de présentations interactives sur le Web.

Le standard MPEG4 vidéo a déjà été utilisé pour encoder les séquences vidéo issues d'un caméra portable. Ce type d'application prend de plus en plus d'importance grâce au transfert rapide et facile vers le Web et pourrait utiliser le mode en texture fixe pour la capture d'images fixes. Pour le marché des jeux, le standard vidéo MPEG4, le codage de textures fixes, l'interactivité, la placage de textures 3D d'images fixes et l'enrichissement de séquences vidéo devrait enrichir les expériences des joueurs.

Les textures naturelles, les images et la vidéo Les outils servant à représenter les objets visuels naturels en MPEG4 fournissent une technologie standardisée pour le stockage, la transmission et la manipulation de textures, d'images et de données vidéo pour les environnements multimédia. Ces outils permettent de décoder et de représenter des unités atomiques d'images ou de

vidéos appelées objets vidéo (Vidéo Objects (VO) en anglais). Un exemple de VO peut être une personne en train de parler (sans le fond) qui peut alors être composé avec d'autres objets audiovisuels pour définir une scène.

Pour atteindre ce but MPEG4 propose des solutions et des algorithmes, regroupant la plupart des fonctionnalités demandées comme pour :

- la compression des images et des vidéos,
- la compression des textures pour les maillages 2D et 3D,
- la compression des maillages 2D implicites,
- la compression des champs d'animation géométrique des maillages,
- l'accès aléatoire de tous types de VO,
- l'extension des fonctionnalités de manipulation des images et des séquences vidéo,
- le codage des vidéos et des images basé sur le contenu,
- le redimensionnement des objets basé sur le contenu,
- le redimensionnement spatial, temporel et qualitatif,
- la robustesse et la résistance aux erreurs quel que soit l'environnement.

Les objets synthétiques Les objets synthétiques englobent une importante partie de l'imagerie par ordinateur. Dans la suite, nous décrirons les objets synthétiques visuels suivant :

- Descriptions paramétriques de :
  - l'animation des champs du visage et du corps,
  - le codage dynamique et statique du maillage avec les textures,
- le codage des textures suivant les vues.

Animation du visage L'objet «animation faciale» peut être utilisé pour afficher un visage animé. La forme, la texture et l'expression du visage sont contrôlées par des p aramètres de définition faciale (FDP pour Facial Definition Parameter en anglais) et/ou des paramètres d'animation faciale (FAP pour Facial Animation Parameters en anglais). A l'initialisation, l'objet visage est un visage générique avec une expression neutre. Il peut recevoir des paramètres

d'animation d'un flux de données tels des expressions, un texte à dire, etc. Des paramètres de définition peuvent aussi être reçus qui changeront l'apparence du visage pour lui donner une apparence nouvelle avec sa forme et sa texture propre. Un modèle complet de visage peut être chargé via le FDP.

Animation du corps La technologie d'animation du corps suivra directement de celle du visage. Les FDP et FAP sont remplacés par des BDP (Body Definition Parameter) et BAP (Body Animation Parameters).

Animation des maillages 2D Le maillage 2D est une partition d'un espace 2D par des polygones eux-mêmes référencés par une liste de nœuds. La norme MPEG4 utilise uniquement le type de maillage triangulaire, longtemps utilisé pour la représentation d'objets 3D. La modélisation par maillage triangulaire peut être considérée comme la projection d'un maillage 3D sur une image plane. La figure 12 en présente un exemple.

MPEG4 utilise un maillage dynamique triangulaire pour conserver la facilité de manipulation et les multiples fonctionnalités qu'offre cette solution pour les objets 3D comme :

- pour la manipulation d'objet vidéo : améliorer le réalisme des scènes, modifier ou remplacer des objets, rendre plus robuste l'interpolation spatio-temporelle lors de la reconstruction des images (en cas de pertes d'information) ;
- pour la compression, le maillage permet d'augmenter le taux de compression avec un faible taux d'erreur.

Echelonnage en fonction des vues En fonction de la façon dont on regarde une scène, toutes les informations ne sont pas nécessaires. L'échelonnage permet de sélectionner uniquement la partie utile de l'information, et donc de transférer une masse d'information considérablement réduite entre la base de données et l'utilisateur, données qui seront traitées sous cette forme réduite au codage et au décodage(compression). Cette méthode est de plus applicable aussi bien avec les ondelettes qu'avec le codeur DCT.

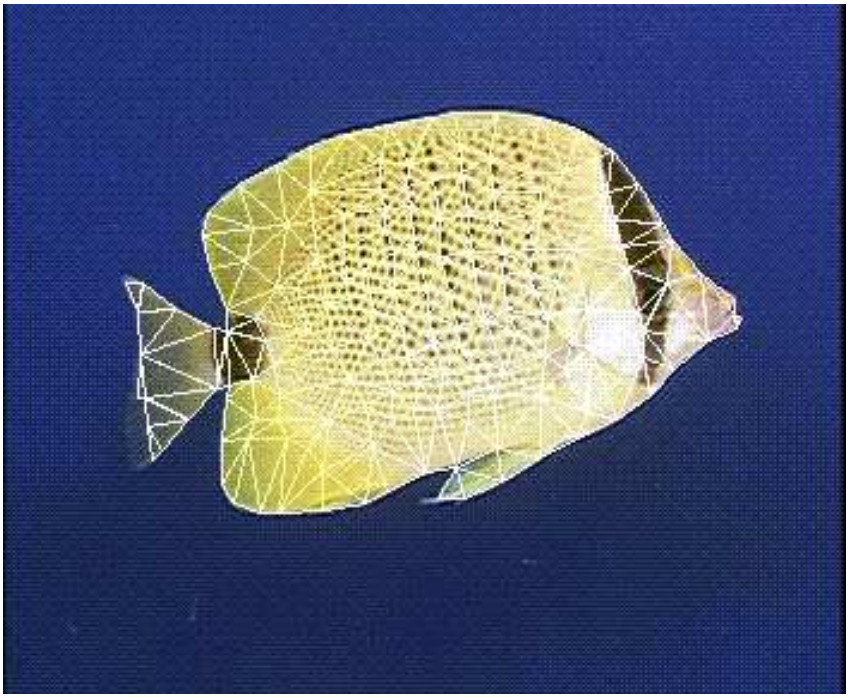

FIG. 12: Exemple de maillage

Structure des outils pour représenter des vidéos naturelles MPEG4 supporte les fonctionnalités déjà fournies par MPEG1 et MPEG2 : la compression des images traditionnelles rectangulaires de différents formats, la fréquence des images, la profondeur des pixels, le taux de transfert, et les possibilités de redimensionnement spatiaux, temporels et qualitatifs. MPEG4 fournit des algorithmes et outils pour des applications à très faible débit (VLBV : Very Low Bit-rate Video, entre 5 et 64kb/s) pour des séquences d'images de faible résolution avec peu d'images par secondes (jusqu'à 15 images/s). Les fonctionnalités principales disponibles pour le noyau VLBV sont :

- codages des séquences d'images rectangulaires avec un fort taux de compression, une grande résistance aux erreurs, une faible latence et une faible complexité pour les applications temps-réel,
- l'accès aléatoire, l'avance et le retour rapides pour les applications de stockage et d'accès à des bases de données multimédia.

Ces outils et algorithmes prévus pour de faibles débits sont aussi efficaces à hauts débits (HBV : de 64kb/s à 10Mb/s) avec une qualité de rendu proche de celui de la TV digitale.

Le codage basé sur le contenu permet de coder et décoder séparément les différents "objets vidéo" (VO) d'une scène. Il permet ainsi une gestion simplifiée de l'interactivité (manipulation et représentation des objets vidéo) et un mélange aisé entre objets naturels et objets synthétiques (comme par exemple une scène avec un fond virtuel avec des personnages réels).

Schéma de codage des objets vidéo et des images La figure 13 présente le schéma de codage des images et vidéos MPEG4, qui permet de traiter les traditionnelles images rectangulaires aussi bien que les formes arbitraires (shape) d'une séquence.

La structure de base du codage utilise le codage de forme (pour les images de forme arbitraire), la compensation de mouvement ou encore le codage de texture par DCT.

Un des avantages les plus intéressants du codage basé sur le contenu est l'amélioration du taux de compression par l'utilisation d'une prédiction de mouvement approprié a chaque objet de la scène. Plusieurs techniques de pré-

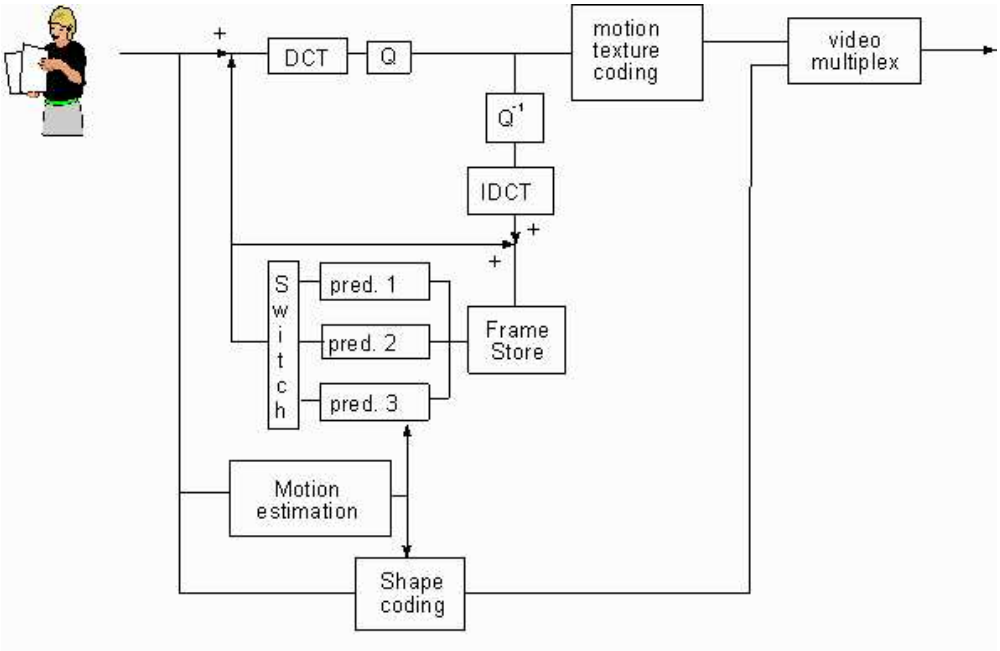

FIG. 13: Schéma de codage des vidéos naturelles en MPEG4

diction de mouvement peuvent être utilisées pour permettre un codage efficace et flexible :

- la prédiction et la compensation de mouvement classique à base de blocs 8 ou 16,
- la compensation de mouvement basée sur l'utilisation de «sprite» (esprit ou farfadet en français !). Un sprite est une image fixe décrivant un fond panoramique. A chaque image successive d'une séquence, seuls 8 paramètres décrivant les mouvements de la caméra sont codés et transmis pour reconstruire l'objet. Ces paramètres représentent la transformation affine du sprite transmis dans la première image.

La figure14 illustre le principe d'utilisation des sprites. On suppose pouvoir extraire le joueur de tennis du reste de l'image par segmentation avant le codage. Le fond de l'image est transmis au receveur uniquement lors de l'émission de la première image et stocké localement chez le receveur. Pour toutes les images suivantes, seules les paramètres des mouvements de la caméra seront transmis ce qui permet au receveur de reconstruire l'image d'origine. L'objet au premier plan (le joueur) est transmis pour chaque image comme un objet de forme arbitraire. Le receveur compose les deux images pour rendre l'image d'origine. Pour les liaisons à faible débit, il est possible de transmettre le sprite par bouts successifs au fur et à mesure du rendu de la séquence.

Le codage adaptable des objets vidéo MPEG4 permet le codage des images et des vidéos de manière adaptée aussi bien dans l'espace que dans le temps pour des objets rectangulaires que des objets de formes quelconques. L'adaptabilité se réfère ici à la possibilité de ne décoder qu'une partie du flot de données et de reconstruire les images ou les séquences d'images avec :

- une complexité de décodage réduite et, donc, une qualité réduite,
- une résolution spatiale réduite,
- une résolution temporelle réduite,
- une qualité réduite à résolution spatiale et temporelle égales.

Cette fonctionnalité permet un codage adapté à de faibles débits quand les images ou vidéos sont envoyées sur un réseau ou encore d'adapter la qualité aux possibilités du receveur (faible CPU, résolution réduite du terminal).

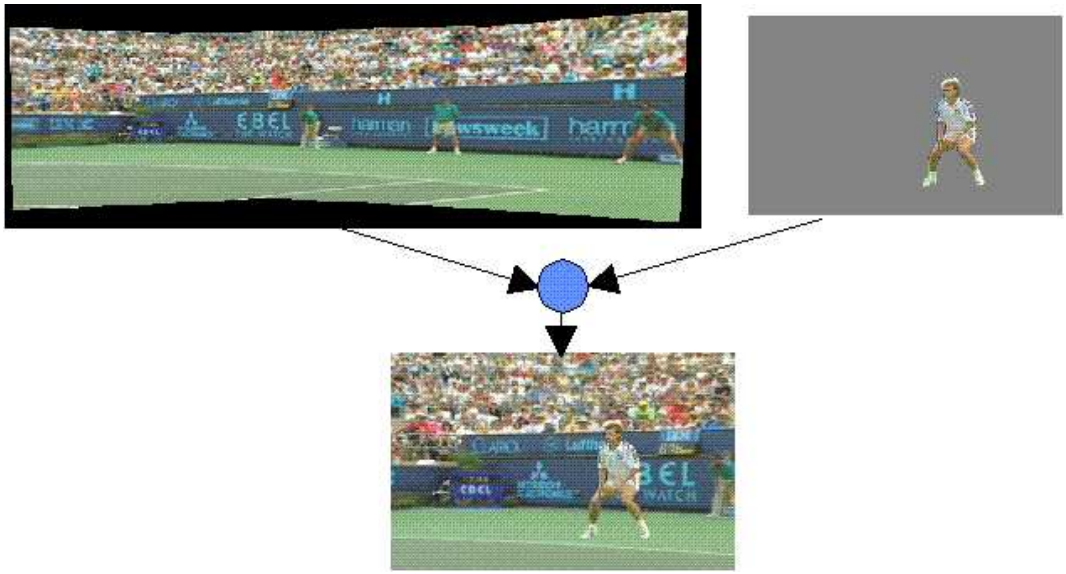

FIG. 14: Exemple de codage basé sur le contenu de vidéos naturelles en MPEG4

Pour les image fixes, la norme a fixé 11 niveaux de granularité et une adaptation de qualité jusqu'au bit. Pour les vidéo, 3 niveaux de granularité ont été définis mais le nombre final devrait être de 9.

6 Les principaux codecs

La liste des codecs actuellement disponibles est longue et les évolutions dans ce domaine sont très rapides, dépendant du rythme des rachats et faillites des différentes entreprises. Ils varient par les méthodes de compression utilisées et les platefomes qui les utilisent. Des méthodes de compression standard ont été définies tant pour le visionnage de vidéo (tels MPEG-1, MPEG-2 et MPEG-4) que pour la vidéoconférence (tels H.261 ou H.263) ou la téléphonie (comme H.323). Pour une liste presque exhaustive des codecs actuels, le lecteur intéressé pourra se référer à http ://www.terran.com/CodecCentral/.

6.1 Cinepak

Ce codec a initialement été conçu pour le `Quick Time` de `Apple` afin de visionner de petites vidéos issues de CD-ROM à simple vitesse (1x). Il a par la suite été porté sous `Windows`, sur certaines consoles de jeux et a fait l'objet d'implémentations matérielles sur des cartes destinées à des PC.

C'est un codec avec perte, orienté bloc, utilisant la quantification vectorielle. La compression est à la fois spatiale et temporelle. Pour avoir une qualité d'image acceptable, le débit doit être supérieur à 30 kb/s. Le temps de compression est important.

Ce codec est supporté par les architectures QuickTime et Video pour Windows.

Un des avantages principaux de ce codec est de ne nécessiter que très peu de ressources CPU. Avec ses évolutions successives, ce codec a été amélioré afin de prendre en compte des débits plus élevés (et plus bas pour le Web) et des films plus longs. S'il ne supporte cependant pas la comparaison avec les tout nouveaux codecs, il reste un bon choix pour distribuer des vidéos qui pourront être visionnées sur le plus grand nombre de machines. Étant supporté à la fois par QuickTime et Video pour Windows, c'est une bonne solution pour distribuer des vidéos multi-plateforme (même si un "transcodage" est nécessaire).

En revanche, la qualité vidéo est inférieure à celle de nombreux autres codecs pour les mêmes bandes passantes en particulier pour les bandes passantes du Web (en dessous de 30 ko/s). Cinepak compresse au moins en 10 : 1, il présente donc moins d'intérêt pour du matériel permettant des bandes passantes plus élevées (CD-ROM 4x et plus).

6.2 Indeo 3.2

Ce codec a été développé pour Windows par Intel dans les années 80 et portait initialement le nom de "Real Time Video 2.1" (ou RT2.1). Ce codec est très proche de Cinepak. Il est adapté au débit des CR-ROM. Le temps de compression sont honnêtes et le codec a été porté sur une grande variété de machines. C'est le prédécesseur du codec Intel Video Interactive (IVI) mais il utilise un algorithme de compression totalement différent.

La compression est une compression avec perte, orientée bloc, utilisant la quantification vectorielle. La compression est à la fois spatiale et temporelle.

Ce codec est supporté par l'architecture QuickTime pour Mac, mais pas par QuickTime pour Windows qui supporte le successeur d'Indeo 3.2, Indeo Video Interactive, qui n'est pas compatible avec Indeo 3.2... Ce codec est supporté par l'architecture Video pour Windows.

La compression avec ce codec est 30% plus rapide que pour Cinepak. Indeo 3.2 est parfois supérieur en terme de qualité et de taux de compression à Cinepak pour des vidéos comme les journaux télévisés où la plupart de l'image de fond est statique.

Du fait de l'incompatibilité avec QuickTime pour Windows, ce codec n'est pas adapté à la production de vidéos multi-plateformes. Quelques problèmes avec les couleurs peuvent parfois apparaître. Un bug connu empêche ce codec de fonctionner correctement avec des vidéos plus hautes que larges.

6.3 RealVideo G2 SVT

Ce codec a été développé par la société `Real Networks`. C'est un codec propriétaire. Peu d'informations sont donc disponibles sur son fonctionnement interne. Il est utilisé par le visionneur de la même société, le RealSystem G2, disponible sur PC et sur Mac. Ce codec utilise la technologie SVT (pour Scalable Video Technology) qui lui permet de s'adapter à la vitesse de connexion en temps réel, si par exemple le débit chute en cours de visualisation. Dans ce cas, le serveur dégrade la qualité des images envoyées. Ce choix est préférable au classique saut des images non arrivées dans les délais. Cette possibilité est rendue aisée par l'utilisation d'un codage par ondelettes qui encode un signal comme une série de raffinements. Ce codec est dérivé du codec Indeo Video Interactive de Intel. Ce codec effectue une compression temporelle sur les vidéos. Accessoirement, ce codec reconnaît le standard SMIL (Standardized Multimedia Integration Language, voir `http://www.w3c.org/AudioVideo/`).

Les qualités généralement reconnues de ce codec sont :

- la possibilité d'encoder une vidéo pour différents débits et d'envoyer automatiquement la bonne vidéo en fonction de la vitesse de la connexion utilisateur

- Le codage à débit variable qui permet d'optimiser l'utilisation de la bande passante en fonction de la complexité des images codées,
- une très bonne qualité pour les débits faibles ou moyens.

Les défauts sont :

- ce codec n'est pas compatible avec le précédent du même constructeur,
- il faut un ordinateur puissant pour obtenir une qualité optimale.

6.4 M-JPEG

M-JPEG, aussi appelé Motion JPEG, est une adaptation à la vidéo de la norme JPEG issue du Joint Photographic Expert Group qui concerne les images fixes. Avec cette norme, une vidéo est simplement considérée comme une succession d'images fixes au format JPEG. Il y a donc compression intra-image puisque la norme JPEG compresse l'image considérée, mais pas de compression inter-image utilisant la redondance entre images successives. Cette norme est donc idéale pour l'édition vidéo : une suppression aléatoire ne risque pas de faire perde une image clé. JPEG a été une des premières méthode de compression d'images à avoir des implémentations matérielles, permettant ainsi une compression en temps réel des vidéos M-JPEG. C'est sans doute une des raisons pour lesquelles cette norme a été adoptée par de nombreux systèmes de montage vidéo commerciaux tels Avid Systems, Radius VideoVision ou encore ImMIX VideoCube.

La compression JPEG est une compression avec perte par bloc. Elle utilise la transformée en cosinus discrète.

Ce codec est supporté par les architectures QuickTime et Video pour Windows.

La qualité des images est généralement du niveau des images télévisées. Ce codec se comporte très bien pour des images de haute résolution, des photographies et des images en nuances de gris. Il est particulièrement adapté aux diaporamas.

Même si JPEG est un standard, la manière dont les différentes implémentations l'utilisent diffère. Différents systèmes répondant au standard M-JPEG ne sont donc pas forcément compatibles. Par exemple, les systèmes de la société Radius utilisent un taux de compression sur les images variables afin de maintenir un débit de sortie constant depuis les disques durs.

Ce standard produisant de très gros fichiers, il est peu utilisé pour la distribution de vidéo au grand public. Étant donnée la lourdeur de la charge de décompression, ce codec n'est pas adapté à la diffusion de vidéo sur CD-ROM 1x ou plus sans avoir recours à du matériel spécialisé (le CPU serait incapable de tenir la cadence). Si un fichier est décompressé puis compressé plusieurs fois, les pertes induites par la compression JPEG cumulées peuvent finir par créer des artefacts gênants.

6.5 MPEG1

Les codecs MPEG sont les seuls qui peuvent revendiquer être de vrais standards. Ils sont basés sur le travail du Motion Picture Expert Group, un groupe de travail de l'Organisation de standardisation international (OSI ou ISO pour les anglophones). Le standard MPEG-1 définit une méthode de compression pour les données audiovisuelles adaptée à la bande passante des lecteurs de CD-ROM.

MPEG-1 utilise une méthode de compression par bloc avec perte basée sur la DCT. La compression est à la fois spatiale et temporelle. La méthode de compression temporelle est légèrement différente de la méthode classique. Classiquement, la redondance temporelle est exploitée en ne stockant pour une image donnée que sa différence avec l'image précédente. Des images de référence sont insérées de manière périodique. Elles sont appelées image clé ou image I (key frame ou I-frame) dans la norme. MPEG définit deux types d'images intermédiaires. Les premières sont appelées image P (pour Past) qui sont basées sur une image de référence précédente. Les secondes sont les images B (pour bidirectionnelles) qui sont basées sur la différence avec les images du passé et du futur. Ces images sont très compressées puisqu'elles sont basées sur la redondance d'informations contenues dans deux autres images.

Cette norme a été définie au départ pour une résolution de 352 pixels. Cette résolution peut être doublée verticalement et horizontalement en dupliquant les pixels, fournissant des images à forte granularité. Le standard a cependant été étendu pour permettre la gestion d'une résolution de 640.

La compression et la décompression ont initialement été prévues pour utiliser du matériel dédié. La compression est très lourde. Il est possible de visua-

liser des vidéos MPEG1 par logiciel avec l'apparition de processeurs de plus en plus puissants.

Du fait de la perte d'information lors de la compression, MPEG1 est essentiellement un format de diffusion et n'est pas adapté au montage vidéo.

La plupart des architectures supportent ce codec que ce soit QuickTime, le Media Player ou le RealPlayer.

La qualité audio et vidéo est excellente pour des débit issus de CD-ROM 1x ou 2x (donc pas vraiment pour le Web). C'est devenu un standard largement diffusé.

6.6 MPEG2

Le standard MPEG-2 définit une méthode de compression pour les données audiovisuelles initialement adaptée à des bandes passantes élevées. Cette norme améliore la résolution vidéo, la qualité du son et le nombre d'images par secondes de la norme MPEG1. Le débit cible est compris entre 4 et 15Mb/s pour une vidéo de qualité télévisuelle. MPEG-2 est le standard choisi pour les vidéos diffusées sur DVD-Vidéo. Ce standard n'est pas adapté à la diffusion sur le Web étant donné les débits nécessaires.

La décompression est une tâche complexe et un dispositif matériel est nécessaire pour obtenir une restitution de bonne qualité. La compression en temps réel est largement atteinte grâce à des dispositifs matériel *ad hoc*.

Pour les mêmes raisons que MPEG1, MPEG2 n'est pas très adapté au montage vidéo.

La plupart des architectures pour le Web ne supportent pas ce codec mais de nombreux visionneurs spécifiques sont disponibles. Un des inconvénients de cette norme que la plupart des machines vendues avant 1998 sont incapables de visualiser de telles vidéos.

6.7 Windows Media Video

Le `Windows Media Video` développé par `Microsoft` est le codec principal du visionneur du même constructeur, le `Windows Media Player`. C'est le seul codec actuellement à revendiquer la compatibilité avec le standard MPEG-4.

C'est un des plus récents et des meilleurs codecs actuels pour la vidéo distribuée par flot (streaming en anglais).

Les qualités généralement reconnues de ce codec sont l'excellente qualité des vidéos distribuées par flot sur le Web, des extensions propriétaires qui permettrait une meilleure qualité que MPEG4.

Les défauts sont la nécessité d'avoir une machine puissante pour visualiser des vidéos à haut débit et la non disponibilité (actuelle) sur Macintosh.

7 Conclusion

Conscients de ne pouvoir prétendre à l'exhaustivité, nous avons exposé dans cette étude les principales techniques de compression mises en œuvre pour les grands standards actuels en matière de diffusion de vidéo. Le standard JPEG utilisé pour les images fixes et le standard MPEG2 adopté pour les DVD sont décrits. La norme MPEG4, non encore totalement finalisée à ce jour, est présentée. La partie vidéo est décrite en détails.

Concernant le choix du codec à utiliser pour distribuer une vidéo avec une qualité télévisuelle, que ce soit pour MPEG2, MPEG4, ou un système propriétaire le problème n'est pas que technique. Des problèmes de droits d'auteur liés aux licences peuvent se poser, et donc aussi des problèmes de coûts.

MPEG2 nécessite soit un ordinateur puissant, soit une carte vidéo capable d'aider à la visualisation en temps réel mais la plupart des ordinateurs modernes répondent à ces critères. Cependant les débits nécessaires à une qualité broadcast semblent élevés (> 512 kb/s) pour une diffusion en temps réel sur Internet. Des travaux récents sont cependant en cours fournissant des outils libres de droits pour la visualisation de vidéos au format MPEG2. Ces outils ont pour buts de réduire les bandes passantes et les ressources CPU nécessaires (voir `http://heroine.linuxave.net/` et `http://www.linuxvideo.org/` par exemple).

MPEG4 n'est pas encore finalisée, les visionneurs sont encore trop rares et propriétaires (Microsoft).

Restent les solutions propriétaires comme les offres de RealVideo ou QuickTime Ces solutions obligent à être tributaire d'un constructeur particulier, ne permettent pas de modifier les sources des outils et ont un coût non négligeable.

Références

[1] Moving Pictures Expert Group (ISO/IEC JTC1/SC29/WG11). MPEG-4 Intellectual Property Management & Protection (IPMP Overview & Application Document, Dec 1998.

[2] Moving Pictures Expert Group (ISO/IEC JTC1/SC29/WG11). Report of the formal verification tests on MPEG-4 Video Error Resilience, Dec 1998.

[3] Moving Pictures Expert Group (ISO/IEC JTC1/SC29/WG11). Overview of the MPEG-4 Standard (Beijing version), July 2000.

[4] Mark Nelson. *La compression de données*. DUNOD, 1993.

[5] Henning Schulzrinne, S. Casner, R. Frederik, and V. Jacobson. RTP : A Transport Protocol for Real-Time Applications. Internet Draft, RFC 1889, `http://www.cis.ohio-state.edu/htbin/rfc/rfc1889.html`, March 1996.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399